Final Report for Period: 06/2007 - 05/2008
Submitted on: 08/12/2008
Principal Investigator: Altunbasak, Yucel .
Award ID: 0514903
Organization: GA Tech Res Corp - GIT
Submitted By:
Altunbasak, Yucel - Principal Investigator
Title:
Understanding the Book of Life: Bayesian Protein Secondary Structure Analysis and its Application to Protein Function Prediction

Project Participants

Senior Personnel

Name: Altunbasak, Yucel
Worked for more than 160 Hours: Yes
Contribution to Project:

Post-doc

Graduate Student

Name: Aydin, Zafer
Worked for more than 160 Hours: Yes
Contribution to Project:

Undergraduate Student

Technician, Programmer

Other Participant

Research Experience for Undergraduates

Organizational Partners

Other Collaborators or Contacts

Please see the report

Activities and Findings

Journal Publications

Z. Aydin and Y. Altunbasak, "Applications of signal processing in genomic research", IEEE Signal Processing Magazine, p. , vol. , ( ). Accepted,

Z. Aydin, Y. Altunbasak, and M. Borodovsky, "Protein secondary structure prediction for a single-sequence using hidden semi-Markov models", BMC Bioinformatics 2006, p. 1, vol. 7:178, (2006). Published,

Z. Aydin and Y. Altunbasak, "Bayesian protein secondary structure prediction with near-optimal segmentations", IEEE Transactions on Signal Processing, p. , vol. , (2006). Submitted,

Books or Other One-time Publications

Web/Internet Site

Other Specific Products

Contributions

Contributions within Discipline:

Contributions to Other Disciplines:

Contributions to Human Resource Development:

Contributions to Resources for Research and Education:

Contributions Beyond Science and Engineering:

Categories for which nothing is reported:

Organizational Partners
Activities and Findings: Any Research and Education Activities
Activities and Findings: Any Findings
Activities and Findings: Any Training and Development
Activities and Findings: Any Outreach Activities
Any Book
Any Web/Internet Site
Any Product
Contributions: To Any within Discipline
Contributions: To Any Other Disciplines
Contributions: To Any Human Resource Development
Contributions: To Any Resources for Research and Education
Contributions: To Any Beyond Science and Engineering

FINAL REPORT

NSF Project Information:

Project Title: Understanding the book of life: Bayesian protein secondary structure analysis and its application to protein function prediction
PI or PIs: Yucel Altunbasak
Institution: Georgia Institute of Technology
Award Number: 0514903

The research activities funded under the grant 0514903 resulted in 3 journal papers and 6 conference papers. In the following, we list all the publications/products as a result of this research activity:

Refereed Journal Publications:

1) Z. Aydin, Y. Altunbasak, and M. Borodovsky, "Protein secondary structure prediction for a single-sequence using hidden semi-Markov models," *BMC Bioinformatics 2006*, 7:178 (30 Mar 2006).
2) Z. Aydin and Y. Altunbasak, "A signal processing application in genomic research: protein secondary structure prediction," *IEEE Signal Processing Magazine*, volume 23, issue 4, pp. 128-131, July 2006.
3) Z. Aydin, Hakan Erdogan, and Y. Altunbasak, "Bayesian protein secondary structure prediction with near-optimal segmentations," *IEEE Transactions on Signal Processing*, volume 55, issue 7, pp. 3512–3525, July 2007.

Refereed Conference Publications:

1) Z. Aydin, Y. Altunbasak, and M. Borodovsky, "Protein Secondary Structure Prediction with semi-Hidden Markov Models," *IEEE Int. Conf. on Acoustics Speech and Signal Processing*, vol. 5, pp. 577-80, Montreal, CA, May 2004.
2) Z. Aydin, Y. Altunbasak and M. Borodovsky, "Protein secondary structure prediction with semi Markov HMMs," in *Proceedings of the 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 4, pp. 2964-7, San Francisco, CA, September 2004.
3) Z. Aydin, T. Akgun, and Y. Altunbasak, "A modified stack decoder for protein secondary structure prediction," *IEEE Int. Conf. on Acoustics Speech and Signal Processing*, vol. 4, pp. 737-40, Philadelphia, PA, USA, March 2005.
4) Z. Aydin, H. Erdogan and Y. Altunbasak, "Protein fold recognition using residue-based alignments of sequence and secondary structure," *IEEE Int. Conf. on Acoustics Speech and Signal Processing*, volume 1, pp. 349-352, Honolulu, HI, April 2007.
5) I. K. Pakatci, Z. Aydin, H. Erdogan and Y. Altunbasak, "Training set reduction methods for protein secondary structure prediction in single sequence condition," *IEEE 15th Signal Processing and Communications Applications*, Eskisehir, Turkey, 11-13 June 2007.
6) Z. Aydin, Y. Altunbasak, I. Pakatci, H. Erdogan, "Training Set Reduction Methods for Protein Secondary Structure Prediction in Single-Sequence Condition," *29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2007 (EMBS 2007), pp. 5025–5028, Lyon, France, 22-26 Aug. 2007.

Curriculum Development:

Developed a graduate course titled "**Bioinformatics and Bio-signal Processing**". Coverage included pairwise alignment of biomolecular sequences, probabilistic models of DNA sequences, Markov chains and hidden Markov models, statistical models of protein domains, multiple sequence alignment methods, protein secondary structure estimation, and building phylogenetic trees.

ANNUAL PROGRESS REPORT

NSF Project Information:

Project Title: Understanding the book of life: Bayesian protein secondary structure analysis and its application to protein function prediction
PI or PIs: Yucel Altunbasak
Institution: Georgia Institute of Technology
Award Number: 0514903
Report Period: June 2006-May 2007

1. Training Set Reduction Methods for Protein Secondary Structure Prediction in Single-Sequence Condition

Description:

Secondary structure prediction is an invaluable tool in determining the three-dimensional structure and the function of proteins. There are two types of prediction algorithms. A single-sequence prediction algorithm does not use information about other homologous proteins. Such an algorithm should be suitable for a sequence with no similarity to any other protein. Second type of algorithm explicitly uses sequences of homologous proteins, which often have similar structures. The prediction accuracy of such an algorithm is higher than one of a single-sequence algorithm due to incorporation of additional evolutionary information from multiple alignments. The accuracy (sensitivity) of the best current prediction methods is around 80% while this number is below 70% for the algorithms in single-sequence category. The theoretical limit of the protein secondary structure prediction accuracy is estimated to be 88%.

Single-sequence algorithms for protein secondary structure prediction are important because a significant percentage of the proteins identified in genome sequencing projects have no detectable sequence similarity to any known protein. Particularly in sequenced prokaryotic genomes, about a third of the protein coding genes are annotated as encoding hypothetical proteins lacking similarity to any protein with known structure. Also, out of the 25,000 genes believed to be present in the human genome, no more than 40-60% can be assigned a functional role based on similarity to known proteins. From the structure prediction standpoint it is important that two or more hypothetical proteins may bear similarity with each other, in which case it still would be possible to incorporate evolutionary information in a structure prediction algorithm. However, many hypothetical proteins do not have detectable similarity to any protein at all. Such ``orphan'' proteins may represent a sizeable portion of a proteome. For an orphan protein, any method of secondary structure prediction performs as a single-sequence method. Therefore, developing better methods of protein secondary structure prediction from

single-sequence has a definite merit as it helps improving the functional annotation of orphan proteins.

Main Contributions:

One way to improve the performance of a single-sequence algorithm is to perform re-training. In this approach, first, the models used by the algorithm are trained by a representative set of proteins and a secondary structure prediction is computed. Then, using a distance measure, the original training set is refined by removing proteins that are dissimilar to the given protein. This step is followed by the re-estimation of the model parameters and the prediction of the secondary structure. In this project, we developed and compared training set reduction methods that are used to re-train the hidden semi-Markov models employed by our IPSSP algorithm developed earlier. We found that the composition based reduction method has the highest performance compared to the alignment based and the Chou-Fasman based reduction methods. In addition, threshold-based reduction performed better than the reduction technique that selects the first 80% of the dataset proteins.

Results:

In our simulations, we used the EVA set of "sequence-unique" proteins [15] derived from the PDB database. We removed sequences shorter than 30 amino acids and arrived to a set of 2720 proteins. To reduce eight secondary structure states used in the DSSP notation to three, we used the following conversion rule: H, G to H; E, B to E; I, S, T, ' ' to L. To evaluate the performance, we chose the three state-per-residue accuracy (Q_3) as the overall sensitivity measure, which is computed as the total number of correctly predicted amino acids divided by the total number of amino acids in the dataset. From the results shown in Tables 1 and 2, the composition based reduction method is slightly more accurate than the other reduction methods. Compared to the condition with no re-training applied, the composition based reduction improves the secondary structure prediction accuracy by 0.6%. Although the prediction accuracy of the composition based method is close to the accuracy of the alignment based reduction methods, its computational complexity is significantly lower.

Method	Q_3(%)
Composition Based	67.01
Alignment Based (SS)	67.00
Alignment Based (AA+SS)	66.92
Alignment Based (AA)	66.69
Chou-Fasman Based v1	66.65
Chou-Fasman Based v2	66.50
No Re-training	66.59

Table 1: Sensitivity measures of the training set reduction methods. The top 80% of the proteins are classified as similar to the input protein.

Method	Q_3(%)
Composition Based	67.17
Alignment Based (SS)	67.12
Alignment Based (AA+SS)	67.06

Table 2: Sensitivity measures of the training set reduction methods. The dataset proteins are classified as similar to the input protein by a threshold.

Comparing the alignment based reduction methods, the best result is obtained by the method that aligns secondary structures only (Row 2 of Tables 1 and 2). Joint alignments of amino acid sequences and secondary structures did not perform better than secondary structure alignments. This is expected because in single-sequence condition the input protein is not statistically similar to dataset proteins at the amino acid level. Therefore, the discriminative power of the amino acid similarity matrix is weaker than the secondary structure similarity matrix.

Conclusions:

We showed that the training set reduction followed by the re-estimation of the model parameters improves the secondary structure prediction accuracy in single-sequence condition. Among the methods being compared, the composition based reduction technique with thresholding generated the most accurate results and had the lowest computational complexity. This is mainly because of the fact that the composition based reduction does not impose strong constraints, which serves to compensate for the errors made in the initial secondary structure prediction. As a future work, we are planning to optimize the threshold parameter used to construct the reduced dataset. In addition, the methods analyzed can be applied to the second class of prediction algorithms, which utilize evolutionary information in the form of alignment profiles or multiple alignments. In that case, we expect the alignment based method to perform significantly better than the other reduction methods because the accuracy of the initial secondary structure prediction will be comparably higher than the accuracy obtained in the single-sequence condition.

Related conference papers in 2007:

1. Z. Aydin, I. K. Pakatci, H. Erdogan, and Y. Altunbasak, “Training Set Reduction Methods for Protein Secondary Structure Prediction in Single-Sequence Condition”, submitted to 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC ’07).

2. I. K Pakatci, Z. Aydin, H. Erdogan and Y. Altunbasak, ”Training Set Reduction Methods for Single-Sequence Protein Secondary Structure Prediction,” submitted to IEEE Signal Processing and Communication Applications Conference (SIU 2007).

2. Protein Fold Recognition using Residue-Based Alignments of Sequence and Secondary Structure

Description:

Protein structure prediction aims to determine the three-dimensional structure of proteins form their amino acid sequences. Three-dimensional structure prediction (also known as tertiary structure prediction) is vital in many aspects. First, biological functions of proteins are characterized by their tertiary structure. Therefore, accurate prediction of structure will provide information on the functional role of the protein. Second, protein structure prediction is an efficient alternative to expensive and time-consuming experimental methods. Third and most important, protein structure prediction enables us to design novel proteins and drugs, which is a fundamental milestone on the path towards curing diseases.

When a protein does not have similarity (homology) to any known fold, threading or fold recognition methods are used to predict structure. Protein threading or fold recognition refers to a class of computational methods for predicting the structure of a protein from the amino acid sequence. The basic idea is that the target sequence (the protein sequence for which the structure is being predicted) is threaded through the backbone structures of a collection of template proteins (known as the fold library) and a goodness of fit score calculated for each sequence-structure alignment. Protein fold recognition problem can be stated as the problem of assigning a protein of unknown structure (target) to one of the known fold classes (templates) as defined in the SCOP or CATH classification standards. Fold recognition methods can be broadly divided into two types: (1) methods that derive a 1-D profile for each structure in the fold library and align the target sequence to these profiles; (2) methods that consider the full 3-D structure of the protein template. Profile-based methods frequently employ secondary structure, solvent accessibility, and evolutionary information to enhance the accuracy and the quality of the predictions. Second group of method that utilize the 3-D representation, the protein structure is modeled as a set of inter-atomic distances *i.e.*, the distances are calculated between some or all of the atom pairs in the structure. This is a much richer and far more flexible description of the structure, but is much harder to use in calculating an alignment. Methods in the second category greatly benefit from those in the first category.

Main Contributions:

Recent approaches in fold recognition follow two major directions, namely machine learning methods (neural networks and support vector machines) and alignment-based methods. In this work, we present a residue-based alignment method as an alternative to the state-of-the-art SSEA method, originally introduced by Przytycka *et al.* [1], and further modified by McGuffin *et al.* [2]. We show that the power of the SSEA method comes from the length normalization instead of the element alignment technique and a similar performance can be achieved using residue-based alignments of secondary

structures by optimizing gap costs. In addition, the residue-based nature allows us to incorporate amino acid similarity matrices such as BLOSUM into the secondary structure similarity scoring and compute joint alignments, which is not possible with the SSEA method. In that case, the fold recognition performance significantly improves.

Results:

In our simulations, we used two benchmark datasets. The first one is introduced by McGuffin and Jones and is a "difficult" set. The second set is provided by Ding and Dubchak and is relatively easy. In all simulations, we used the sensitivity as the performance measure, which is defined as:

$$Q = \frac{N_c}{N},$$

where N_c is the number of targets with correctly predicted fold classes, and N is the total number of targets evaluated.

We compared the fold recognition accuracies of our method and the SSEA approach. In addition, we evaluated the effect of incorporating amino acid similarity matrix (BLOSUM30) into the residue-based alignments of secondary structure. Tables 3, and 4 show the simulation results on the McGuffin, and Ding and Dubchak sets, respectively. Here, RBSS refers to the residue-based alignments of secondary structure using the secondary structure similarity matrix. The secondary structures were predicted using PSIPRED v2.4. To serve as a reference point, we also computed the alignments using true secondary structure assignments obtained from the PDB. In simulations with the residue-based alignments, the gap opening and gap extension parameters are set to d = -6, and e = -1.2, respectively, which are optimized for the highest recognition accuracy.

Method	Q(%)
SSEA	26.98
RBSS	29.62
BLOSUM30 + RBSS (Predicted SS)	33.73
BLOSUM30 + RBSS (True SS)	35.31

Table 3: Fold recognition accuracy evaluated on the McGuffin set.

Method	Q(%)
SSEA	60.47
RBSS	60.73
BLOSUM30 + RBSS (Predicted SS)	70.68
BLOSUM30 + RBSS (True SS)	75.39

Table 4: Fold recognition accuracy evaluated on the Ding and Dubchak set.

From these results, the residue-based alignment of secondary structures performs comparable or better than the SSEA method. In addition, the incorporation of amino acid similarity scores such as BLOSUM30 brings significant improvements over the secondary structure alignments, which offers a better starting point for more elaborate techniques that employ profile-profile alignments and machine learning methods.

Related conference papers in 2007:

1. Z. Aydin, H. Erdogan and Y. Altunbasak, "Protein Fold Recognition using Residue-Based Alignments of Sequence and Secondary Structure," accepted to IEEE Int. Conf. on Acoustics Speech and Signal Processing (ICASSP '07), 2007.

3. Protein Secondary Structure Prediction with Near-Optimal Segmentations

Description:

Typically, protein secondary structure prediction methods suffer from low accuracy in β-strand predictions, where non-local interactions play a significant role. The β-strand sensitivity of a typical single-sequence prediction method is around 40-50% and that of a method using evolutionary (or homology) information is around 60-70%. There is a considerable need to model such long-range interactions that contribute to the stabilization of a protein molecule.

The proposed hidden semi-Markov model for protein secondary structure prediction has some limitations due to the assumptions made in the model derivation. For instance, it is assumed that the segment likelihood terms are independent from each other. This assumption enables to implement efficient hidden Markov models. However, with this assumption and others inherent in the theory of hidden Markov models, it is not possible to model long-range interactions especially the non-local hydrogen bonds in β-sheet units. More complex dependency models are not feasible due to limitations in the available training data and high computational requirements.

Main Contributions:

To overcome the difficulties in non-local interaction modeling, we propose a two-stage approach. The first step generates a list of best scoring prediction sequences, i.e., *N-best list* that contains the most likely prediction sequence as well as those that are suboptimal under a predefined statistical model. Such a model contains local dependency information and is relatively simple. In the second step, the score of each sequence is updated using a non-local interaction model that utilizes information related to long-range interactions. The final prediction sequence can then be computed using a weighted voting scheme applied to a selected set of top scoring sequences.

To generate suboptimal segmentations, *i.e.*, alternative prediction sequences, we developed two N-best search algorithms that employ hidden semi-Markov models. We developed a Bayesian probability model that characterize the long-range base pairing interaction propensities of the amino acid pairs in β-sheets structures and incorporated the model into the N-best decoding algorithms.

Results:

In the first experiment, we tested the accuracy of the secondary structure prediction method in single-sequence condition. We compared the performances of the Viterbi algorithm (MAP estimator) and the N-Best decoding algorithm that utilizes the non-local interaction model since the N-Best decoder employs the MAP scoring procedure (joint probability of amino acid and secondary structure) to generate suboptimal segmentations. We chose the PDB_SELECT as our dataset and performed leave one out cross-validation. To reduce the computational cost, we restricted our test data to proteins with 2 or 3 beta-strands and eliminated proteins that contain α-helices. The sensitivity results are summarized in Table 5. There is a 5% increase in the beta-strand prediction accuracy as compared to the Viterbi algorithm. However, the loop accuracy decreased, which in turn decreased the overall prediction accuracy.

Sensitivity	Q_3(%)	Q_α(%)	Q_β(%)	Q_L(%)
Viterbi	76.455	--------	26.774	88.097
N-Best decoder with NL interaction model	73.371	--------	31.350	83.217

Table 5: Prediction sensitivity measures, Q_i (%), evaluated on the PDB_SELECT set under the single-sequence condition. N-Best list size is set to 1,000,000. The best scoring 1000 segmentations are combined using a weighted majority voting procedure.

In the second experiment, we analyzed the ranks of the true secondary structure segmentation. For this purpose, we first generated an N-Best list and manually included the true segmentation to the list. Then, we computed the rank of the true segmentation after the score update step. The rank of the true segmentation for proteins with up to 4 β-strands is shown in Table 6.

# Beta-Strands	Rank interval
2	[1-20]
3	[1-500]
4	[1-10000]

Table 6: Ranks of the true secondary structure segmentations.

As depicted in Table 6, the rank of the true segmentation significantly decreases as the number of β-strands in the protein increases. This is because of the fact that for higher number of β-strands, the potential number of β-sheet architectures increases and the mutual signal from the complementary β-strands fades.

All these findings strongly imply that further improvements in secondary structure prediction accuracy are not possible in single-sequence condition. The reason for this behavior is in single-sequence condition, we are restricted to use the frequency of occurrence counts to estimate the base pairing interaction probabilities of the amino acid pairs. These probabilities are uninformative to discriminate true segmentations from the incorrect ones.

Having explored the problem in single-sequence condition, we are currently working on the incorporation of the non-local interaction model into a third generation method that uses multiple alignment profiles. In this setting, we expect the true segmentation to get higher ranks, which is potentially promising for further improvements in the prediction accuracy.

Related journal papers in 2006:

1. Z. Aydin and Y. Altunbasak, "Bayesian protein secondary structure prediction with near-optimal segmentations," accepted to IEEE Transaction on Signal Processing, 2006.
2. Z. Aydin and Y. Altunbasak, "A signal processing application in genomic research: protein secondary structure prediction," IEEE Signal Processing Magazine, vol: 23 Issue: 4, Page(s): 128- 131, July 2006.

ANNUAL PROGRESS REPORT

NSF Project Information:

Project Title: Understanding the book of life: Bayesian protein secondary structure analysis and its application to protein function prediction
PI or PIs: Yucel Altunbasak
Institution: Georgia Institute of Technology
Award Number: 0514903
Report Period: June 2005-June 2006

1. Protein Secondary Structure Prediction for a Single-Sequence using Hidden Semi-Markov Models

Description:

Secondary structure prediction is an invaluable tool in determining the three-dimensional structure and the function of proteins. There are two types of prediction algorithms. A single-sequence prediction algorithm does not use information about other homologous proteins. Such an algorithm should be suitable for a sequence with no similarity to any other protein. Second type of algorithm explicitly uses sequences of homologous proteins, which often have similar structures. The prediction accuracy of such an algorithm is higher than one of a single-sequence algorithm due to incorporation of additional evolutionary information from multiple alignments. The accuracy (sensitivity) of the best current prediction methods is around 80% while this number is below 70% for the algorithms in single-sequence category. The theoretical limit of the protein secondary structure prediction accuracy is estimated to be 88%.

Single-sequence algorithms for protein secondary structure prediction are important because a significant percentage of the proteins identified in genome sequencing projects have no detectable sequence similarity to any known protein. Particularly in sequenced prokaryotic genomes, about a third of the protein coding genes are annotated as encoding hypothetical proteins lacking similarity to any protein with a known function. Also, out of the 25,000 genes believed to be present in the human genome, no more than 40-60% can be assigned a functional role based on similarity to known proteins. From the structure prediction standpoint it is important that two or more hypothetical proteins may bear similarity with each other, in which case it still would be possible to incorporate evolutionary information in a structure prediction algorithm. However, many hypothetical proteins do not have detectable similarity to any protein at all. Such ``orphan'' proteins may represent a sizeable portion of a proteome. For an orphan protein, any method of secondary structure prediction performs as a single-sequence method. Therefore, developing better methods of protein secondary structure prediction from single-sequence has a definite merit as it helps improving the functional annotation of orphan proteins.

Main Contributions:

In this work, we further refine and extend the hidden semi-Markov model (HSMM) initially considered in the BSPSS algorithm. We introduce an improved residue dependency model by considering patterns of statistically significant amino acid correlation at structural segment borders. We also derive models that specialize on different sections of the dependency structure and incorporate them into HSMM. In addition, we implement an iterative training method to refine estimates of HSMM parameters.

Results:

The three-state-per-residue accuracy (sensitivity measure) and other accuracy measures of the new method, IPSSP, are shown to be comparable or better than ones for BSPSS as well as for PSIPRED, tested under the single-sequence condition.

The three-state-per-residue accuracy (Q_3) is defined as:

$$Q_3(\%) = \frac{N_c}{N} \times 100,$$

where N_c is the total number of residues with correctly predicted secondary structure, and N is the total number of observed amino acids. The same measure can also be used for each type of secondary structure, Q_α, Q_β and Q_L:

$$Q_i(\%) = \frac{N_c^i}{N^i} \times 100,$$

where N_c^i is the total number of residues with correctly predicted secondary structure of type i, and N^i is the total number of amino acids observed in conformation of type i.

We first compared the performances of BSPSS and IPSSP on the EVA set. From the results shown in Table 1, there is a 1.9% increase in the overall 3-state prediction accuracy in comparison with the BSPSS method.

Sensitivity	Q_3(%)	Q_α(%)	Q_β(%)	Q_L(%)
BSPSS	68.400	63.203	36.737	82.167
IPSSP	70.300	65.934	45.445	81.280

Table 1: Prediction sensitivity measures, Q_i (%), evaluated on the EVA set under the single-sequence condition. The following conversion rule was applied to reduce the number of states in secondary structure sequences from 8 to 3: H to H, E to E and others to L.

To further verify our results, we compared the performances of the three methods BSPSS, IPSSP and, PSIPRED v2.0 (single sequence version) on 81 CASP6 targets that

are available in PDB (Protein Data Bank). From the results shown in Table 2, and, Table 3, IPSSP is comparable to PSIPRED and is more accurate than BSPSS.

Sensitivity	Q_3(%)	Q_α(%)	Q_β(%)	Q_L(%)
BSPSS	66.541	75.177	41.743	72.696
PSIPRED v2.0	67.680	76.066	52.032	69.028
IPSSP	67.899	74.984	46.087	73.755

Table 2: Prediction sensitivity measures, Q_i (%), evaluated on the CASP6 targets. The following conversion rule was applied to reduce the number of states in secondary structure sequences from 8 to 3: H, G to H, E, B to E and others to L.

MCC	MCC_α	MCC_β	MCC_L
BSPSS	0.5403	0.4354	0.4457
PSIPRED v2.0	0.5465	0.4801	0.4646
IPSSP	0.5657	0.4486	0.4696

Table 3: Matthew's correlation coefficient values, evaluated on the CASP6 targets. The following conversion rule was applied to reduce the number of states in secondary structure sequences from 8 to 3: H, G to H, E, B to E and others to L.

Related journal papers in 2006:

1. Z. Aydin, Y. Altunbasak, and M. Borodovsky, “Protein secondary structure prediction for a single sequence using hidden semi-Markov models,” BMC Bioinformatics, vol. 7, no. 178, 2006.

Related conference papers in 2004:

1. Z. Aydin, Y. Altunbasak, and M. Borodovsky, "Protein secondary structure prediction with semi-hidden Markov models," IEEE Int. Conf. on Acoustics Speech and Signal Processing, vol. 5, pp. 577-80, Montreal, CA, May 2004.

2. Z. Aydin, Y. Altunbasak and M. Borodovsky, "Protein secondary structure prediction with semi Markov HMMs", in Proceedings of the 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, vol. 4, pp. 2964-7, San Francisco, CA, September 2004.

2. Protein Secondary Structure Prediction with Near-Optimal Segmentations

Description:

Typically, protein secondary structure prediction methods suffer from low accuracy in β-strand predictions, where non-local interactions play a significant role. The β-strand

sensitivity of a typical single-sequence prediction method is around 40-50% and that of a method using evolutionary (or homology) information is around 60-70%. There is a considerable need to model such long-range interactions that contribute to the stabilization of a protein molecule.

The proposed hidden semi-Markov model for protein secondary structure prediction has some limitations due to the assumptions made in the model derivation. For instance, it is assumed that the segment likelihood terms are independent from each other. This assumption enables to implement efficient hidden Markov models. However, with this assumption and others inherent in the theory of hidden Markov models, it is not possible to model long-range interactions especially the non-local hydrogen bonds in β-sheet units. More complex dependency models are not feasible due to limitations in the available training data and high computational requirements.

Main Contributions:

To overcome the difficulties in non-local interaction modeling, we propose a two-stage approach. The first step generates a list of best scoring prediction sequences, i.e., *N-best list* that contains the most likely prediction sequence as well as those that are suboptimal under a predefined statistical model. Such a model contains local correlation information and is relatively simple. In the second step, the score of each sequence is updated using a non-local correlation model that utilizes information related to long-range interactions. The final prediction sequence can then be computed using a weighted voting scheme applied to a selected set of top scoring sequences.

To generate suboptimal segmentations, *i.e.*, alternative prediction sequences, we developed two N-best search algorithms. The first one is an A* stack decoder algorithm that extends theories by one symbol at each iteration. The second algorithm locally keeps the end positions of the highest scoring K previous segments and performs backtracking. Both algorithms employ a hidden semi-Markov model.

To incorporate long-range interactions into the current N-best list methods, we are currently developing dependency models that characterize hydrogen bonding propensities of the amino acid pairs in β-strand segments that interact to form β-sheet structures.

Related journal papers in 2006:

3. Z. Aydin and Y. Altunbasak, “Bayesian protein secondary structure prediction with near-optimal segmentations,” submitted to IEEE Transaction on Signal Processing, 2006.
4. Z. Aydin and Y. Altunbasak, “A signal processing application in genomic research: protein secondary structure prediction”, IEEE Signal Processing Magazine, to appear in June 2006.

Related conference papers in 2005:

5. Z. Aydin, T. Akgun, and Y. Altunbasak, "A modified stack decoder for protein secondary structure prediction," IEEE Int. Conf. on Acoustics Speech and Signal Processing, vol. 4, pp. 737-40, Philadelphia, PA, USA, March 2005.